THE LANCET
Planetary Health

Supplementary appendix

This appendix formed part of the original submission and has been peer reviewed. We post it as supplied by the authors.

Supplement to: Santos-Vega M, Lowe R, Anselin L, et al. Quantifying climatic and socioeconomic drivers of urban malaria in Surat, India: a statistical spatiotemporal modelling study. *Lancet Planet Health* 2023; **7:** e985–98.

Table of Contents

Appendix Materials: Quantifying climatic and socio-economic drivers of urban malaria in Surat, India: a statistical spatio-temporal modeling study

Authors

Mauricio Santos-Vega[1,2*], Rachel Lowe [3,4,5], Luc Anselin [6], Vikas Desai [7], Keshav G. Vaishnav [8], Ashish Naik [9], Mercedes Pascual[10**].

Affiliations

1. Departamento de Ciencias Biológicas, Universidad de los Andes, Bogotá, D.C., Colombia
2. Grupo de Investigación en Biología Matemática y Computacional BIOMAC, Universidad de los Andes, Bogotá Colombia.
3. Barcelona Supercomputing Center (BSC), Barcelona, Spain.
4. Catalan Institution for Research and Advanced Studies (ICREA), Barcelona, Spain.
5. Centre on Climate Change & Planetary Health and Centre for Mathematical Modelling of Infectious Diseases, London School of Hygiene & Tropical Medicine, London, UK
6. Center for Spatial Data Science, University of Chicago, Chicago, United States of America.
7. Urban Health and Climate Resilience Center of Excellence (UHCRCE) Surat VBDC, Health Department, Surat Municipal Corporation.
8. Insecticide officer, Vector-Borne Diseases Control Department, Surat Municipal Corporation
9. MOH, Health Department, Surat Municipal Corporation
10. Department of Ecology and Evolution, University of Chicago, Chicago, United States of America.

Appendix 1: Supplementary figures and tables

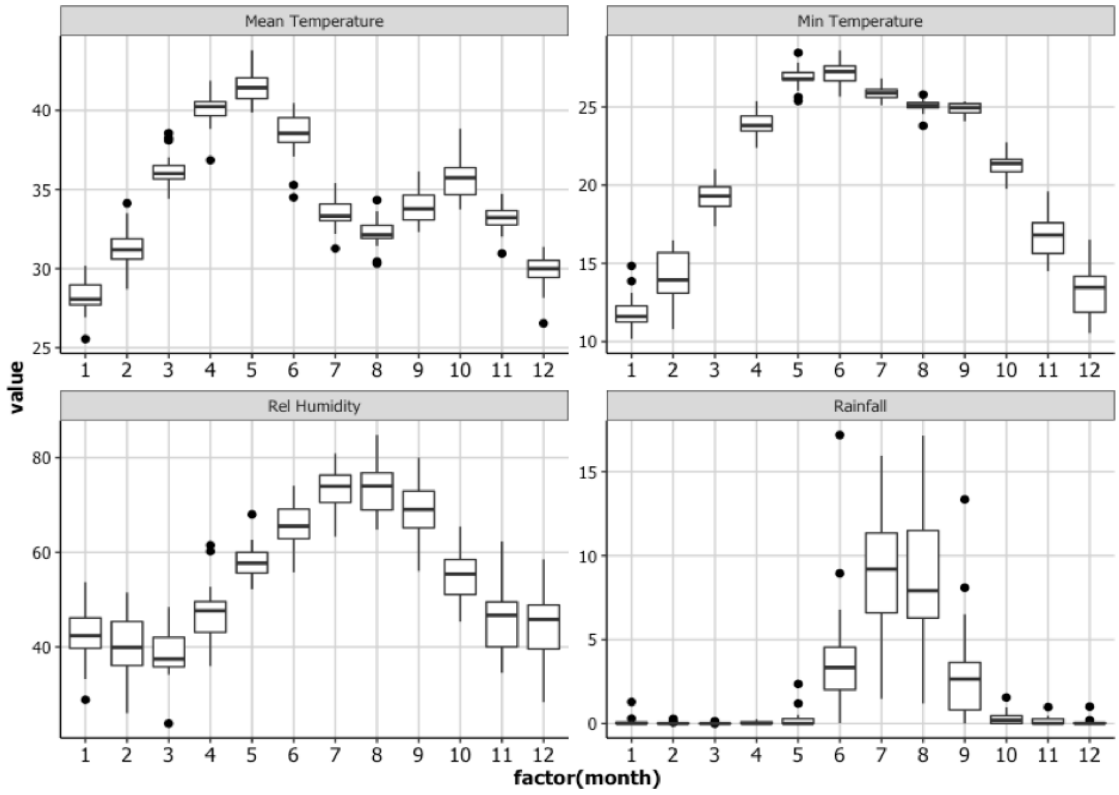

Figure S1. Monthly summaries of (a) mean temperature, (b) minimum temperature, (c) relative humidity, and (d) rainfall in the city of Surat from 2008 to 2015, based on Surat station data from the India Meteorological Department (http://dsp.imdpune.gov.in/)

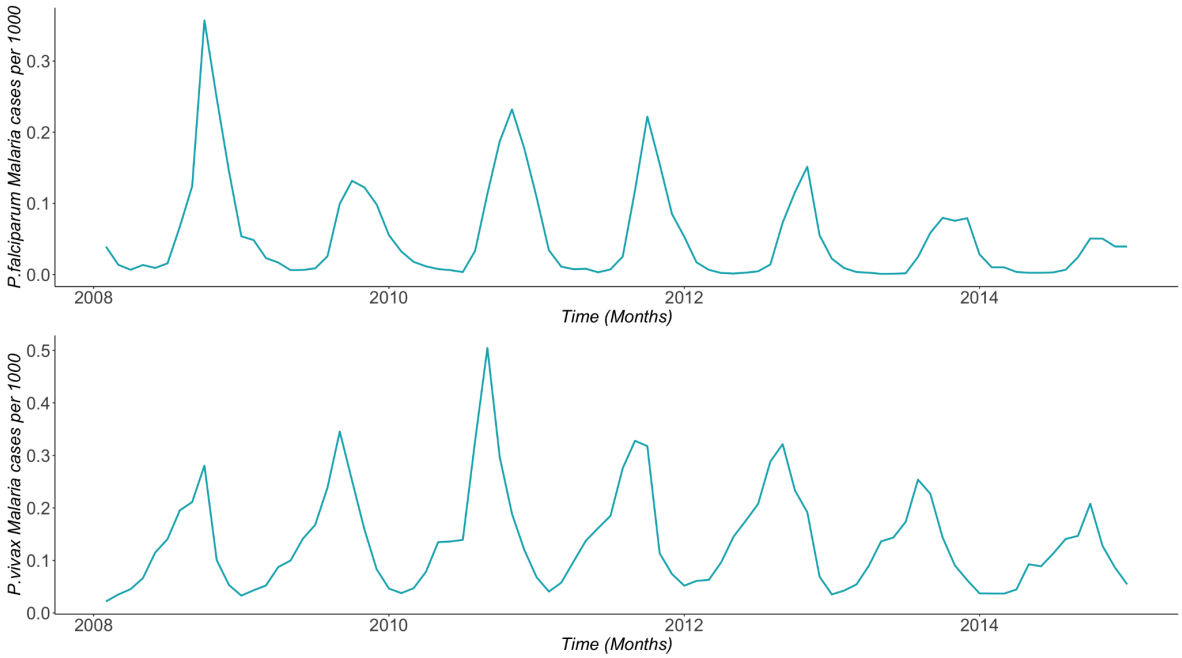

Figure S2. Monthly reported a) falciparum and b) vivax incidence per 1000 people in Surat.

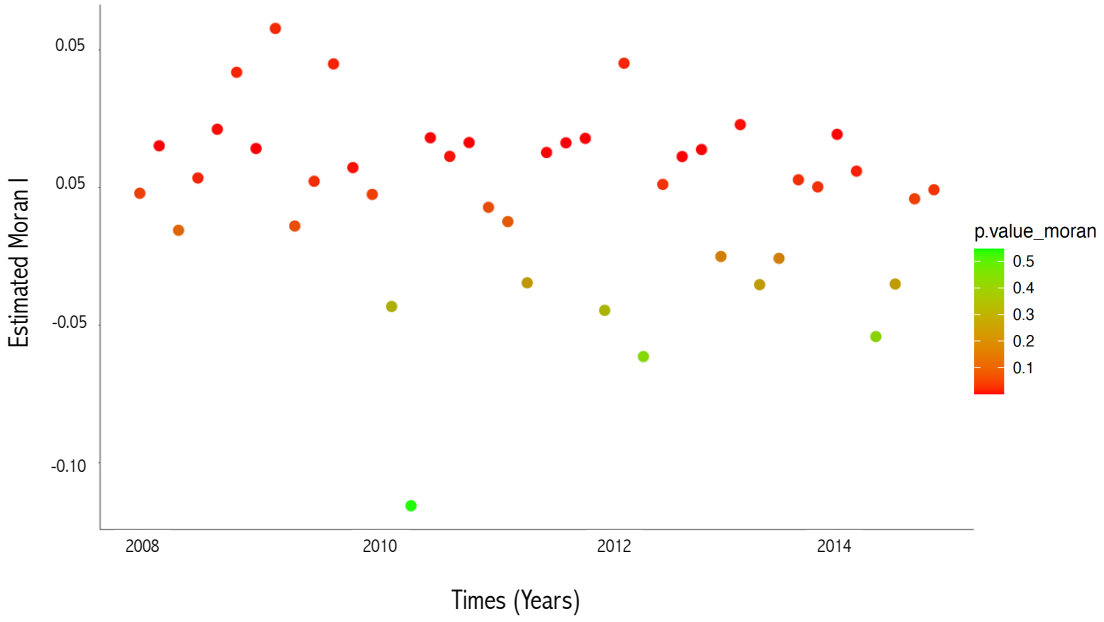

Figure S3. Moran's I autocorrelation index. The index is calculated bimonthly from 2008-2014. Color in the figures from red to green represents the significance of the spatial autocorrelation; red values are associations that are significant at the 0.05 level.

Table S1. Socioeconomic covariates included in the analysis.

Variable	**unit**	**Description**	**Justification**
Total industrial labor	persons	Number of people dedicated to industrial jobs	This variable is a proxy of employment in the city
Total workers	persons	The workers comprise 312 million main workers and 88 million marginal workers (defined as those who did not work for at least 183 days in the preceding 12 months to the census taking)	This variable is a proxy of employment in the city
Density	persons /area	This variable accounts for the number of human hosts per square kilometer in the city.	This variable accounts for unobserved effects in the force of infection, including increasing vector recruitment with human density, as in Romeo-Aznar et al. 2018(59).
Total Income	Rupees/per person	Per capita income in India is the mean income of the people in an economic unit. It is calculated by measuring all sources of income in the aggregate and dividing it by the total population.	We include this variable since we expect it to correlate with disease risk and exposure.
Household area	Sq meters	This is the average house for each unit (ward).	We include this variable since it can be a proxy of the economic level in LMIC
Water stored	Cubic meters.	This variable is considered because the vector breeds in water containers.	Given the vector breeds within houses, the amount of water stored is a relevant covariate.
Total members in the household	persons		This variable can be relevant as related to poverty and local population density.
Distance to river	km	Proximity to rivers	This variable could be correlated with humidity.

Total agriculture labor	person	Amount of people dedicated to agricultural activities.	This variable could be a proxy for mobility to the outer part of the city to work on agricultural jobs.
SC (Scheduled Castes)	persons	officially designated groups of historically disadvantaged people in India.	This is a proxy of vulnerable communities (usually economically deprived) in the city
ST (Scheduled tribes)	persons	Indigenous people officially regarded as socially disadvantaged.	This is a proxy of vulnerable communities (usually economically deprived) in the city
Time to work	Hr/day.	This is the amount of time people commute.	This variable is a proxy for the average commuting time for the people of each unit.
Distance to water sources	meters	This is the distance to the main water bodies in the city.	This distance can be a measure of access to water or the humidity of the environment.
Scarcity	--	This index accounts for the lack of water in some units.	This variable provides an indirect measure for the amount of water that every unit stores in its houses.

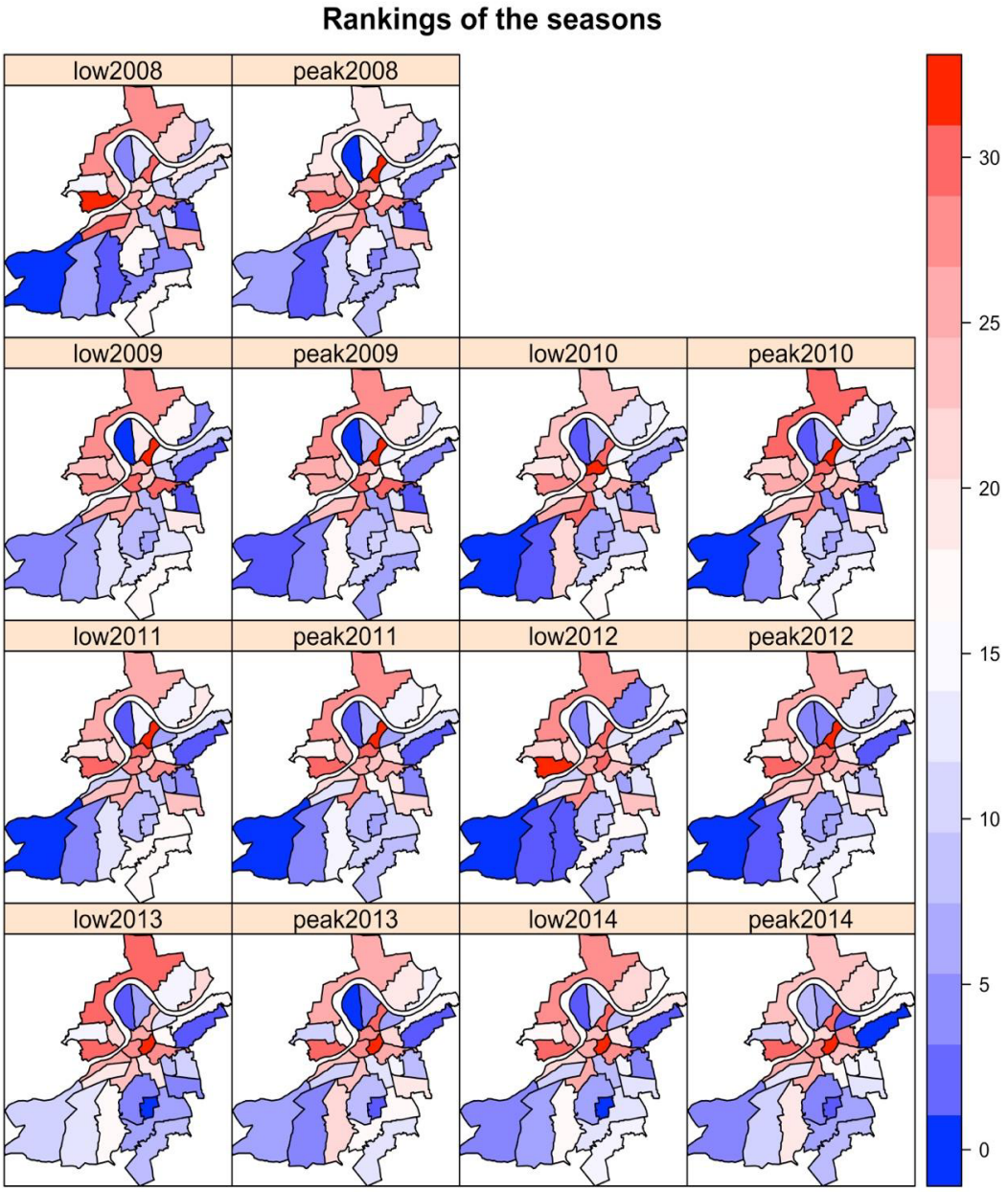

Figure S4. Spatial patterns malaria. The panels show the distribution of malaria rankings based on incidence (from low [blue] to high [red]) for *P. falciparum*.

Table S2. PCA factor loadings of variables (rows) to the different PCs (columns)

Variable	**PC1**	**PC2**	**PC3**
Total industrial labor	0.57380	19.86171	4.52830
Total workers	2.32322	21.77676	0.02798
Population density	1.15284	0.33798	28.57440
Total Income	14.50749	0.47411	7.47029
Household area	7.83117	4.14070	1.35731
Water stored	13.81664	1.85728	1.05474
Total members	12.0220	1.87349	0.00246
Distance	5.81268	0.77895	11.58191
Total agriculture labor	5.15275	1.08217	7.88543
SC	1.41002	23.06431	0.01058
ST	2.97156	20.55982	0.08122
Timework	3.40557	1.84227	24.78332
Distance to water	14.38613	1.23322	2.00411
Scarcity	14.63408	1.11722	0.63794

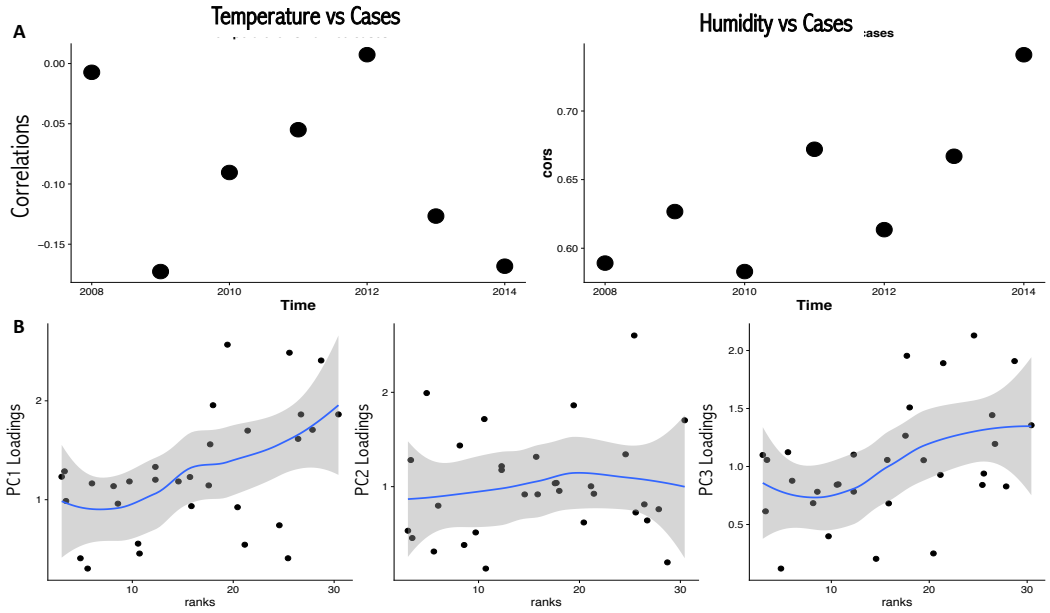

Figure S5. Correlations between cases and drivers. A) Top panels show temporal correlation plots between cases and temperature ($R^2 = 0.21$, $p > 0.01$) and humidity ($R^2 = 0.72$, $p < 0.01$), respectively. B) Bottom panels show correlation plots between the cases and the PCA scores of the three dimensions identified with the PCA in figure 2. These plots specifically show significant associations between cases and PC1 (R^2=0.65, $p < 0.05$) and cases and PC3 (R^2=0.53,$p < 0.05$).

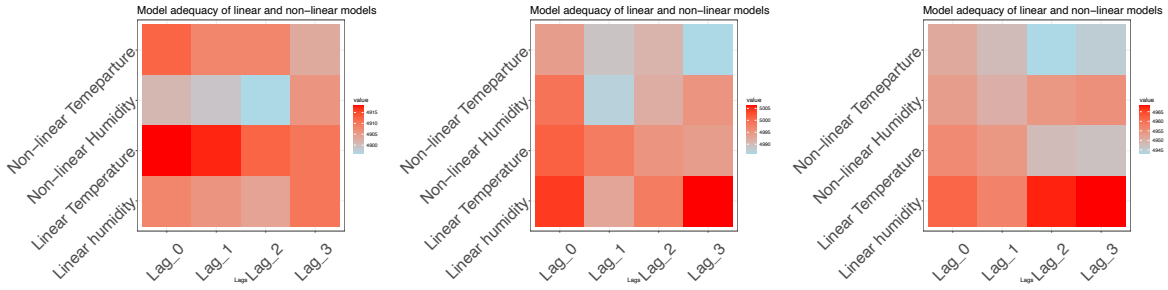

Figure S6. Adequacy of linear and non-linear models for malaria incidence at different lags (0-3). Deviance information criterion (DIC) for spatiotemporal models of monthly *P. falciparum* cases as log(**μ**st). The models include the linear effect of climate and the non-linear effect of climate (temperature and humidity) at the levels of the units.

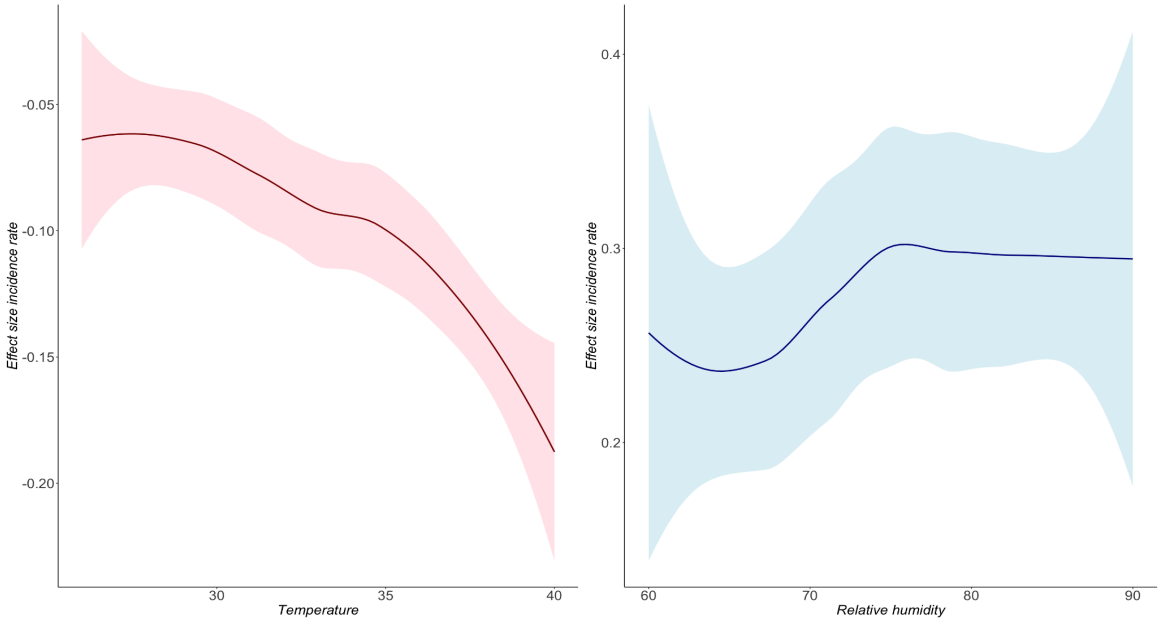

Figure S7. Influence of climate variation on malaria incidence. Effect size of mean temperature (A) and humidity (B) on malaria incidence rate. The best identified model included nonlinear temperature and humidity functions, indicating a nonlinear relationship between transmission and malaria

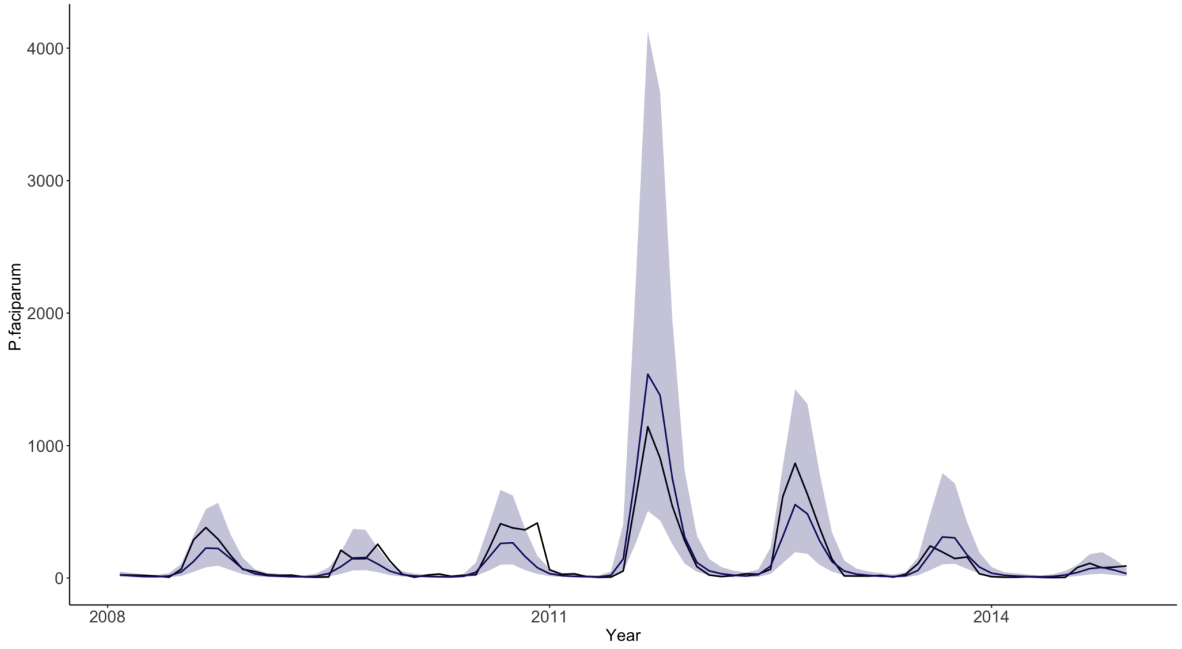

Figure S8. Posterior median incidence. Dark Blue solid curve and 95% credible intervals (Purple gray shaded region) aggregated across all units from 2005 to 2014. The observed incidence is shown in blue.

Table S3. Parameter estimates for the coefficients associated with the respective variables included in the selected model (i.e., the variables explaining the variation in the urban malaria cases). The credible intervals (CI) correspond to the marginal posterior distribution's 2.5 and 97.5% quantiles. Note that the over-dispersion parameter of the negative binomial (i.e., the reciprocal of the scale parameter) has a posterior mean value of 2.519 with a 95% credible interval (CI) of [1.456, 3.243]. Thus, the estimated over-dispersion parameter significantly differs from infinity (the value expected for the Poisson special case of the negative binomial).

covariate	**aggregation**	**median**	**IC_low**	**IC_high**
Non-linear spatial temperature	Units	-0.256	-0.289	-0.226
Non-linear global humidity	Units	0.598	0.502	0.704
Principal component 1	Unit	0.105	0.049	0.223
Principal component 3	Units	0.546	0.369	0.724
Monthly random effect	Units	0.436	0.389	0.511
Yearly random effect	Units	1.225	0.901	1.322
Spatially structured random effect	Units	1.815	1.694	2.321
Spatially unstructured random effect	Units	0.715	0.51	1.234
overdispersion parameter	Units			

Table S4. Predictability evaluation at different levels. The number in the first column the corresponds to the number of units where the model predicts the same quantile as observed. The number in the second column represents the percentage of units that predict the same quantile as observed.

Model	Number of units that predict the same quantile as observed.	Average proportion of units predicting same quantiles as observed
Aggregated	(2)	29%
Intermediate	(18)	57%
Disaggregated	(236)	46%

Table S5. Goodness of fit for the spatio-temporal statistical model at the coarser scale of zones and the finer scale of worker units. Adequacy results include deviance information criterion (DIC), and likelihood-ratio-based R2 for different models of *P. falciparum* cases in Surat. Lower DIC values indicate a better fit and higher R2 indicate a larger fraction of the variability explained.

Model	**Coarser resolution (Zones) DIC**	**Likelihood-ratio based pseudo-R-squared (Zones)**	**Finer resolution (workers) DIC**	**Likelihood-ratio based pseudo-R-squared (Zones**
Baseline: month (e.g. seasonality) and year random effects				
(A) Unstructured + structured spatial random effects, monthly (e.g. seasonality) and yearly random effects	5012	0,221	4971	0,282
Baseline + socio-economic (different combinations of PCA1, 2 and 3)				
(B) Baseline + PC1	5010	0,242	4956	0,351
(C) Baseline + PC1 + PC2	5015	0,207	4959	0,324
(D) Baseline + PC1 +PC2 + PC3	5016	0,199	4960	0,312
(E) Baseline + PC2 + PC3	5021	0,146	4961	0,316
(F) Baseline + PC1 + PC3	5001	0,266	4952	0,368
Baseline + socio-economic + different combinations of climate factors				

(G) Baseline + PC1 + PC3 + temperature linear	4996	0,331	4949	0,387
(H) Baseline + PC1 + PC3 + global temperature non-linear	4989	0,352	4945	0,413
(I) Baseline + PC1 + PC3 + spatial temperature non-linear	4981	0,381	4943	0,424
(J) Baseline + PC1 + PC3 + spatial temperature nonlinear + linear humidity	4978	0,396	4942	0,441

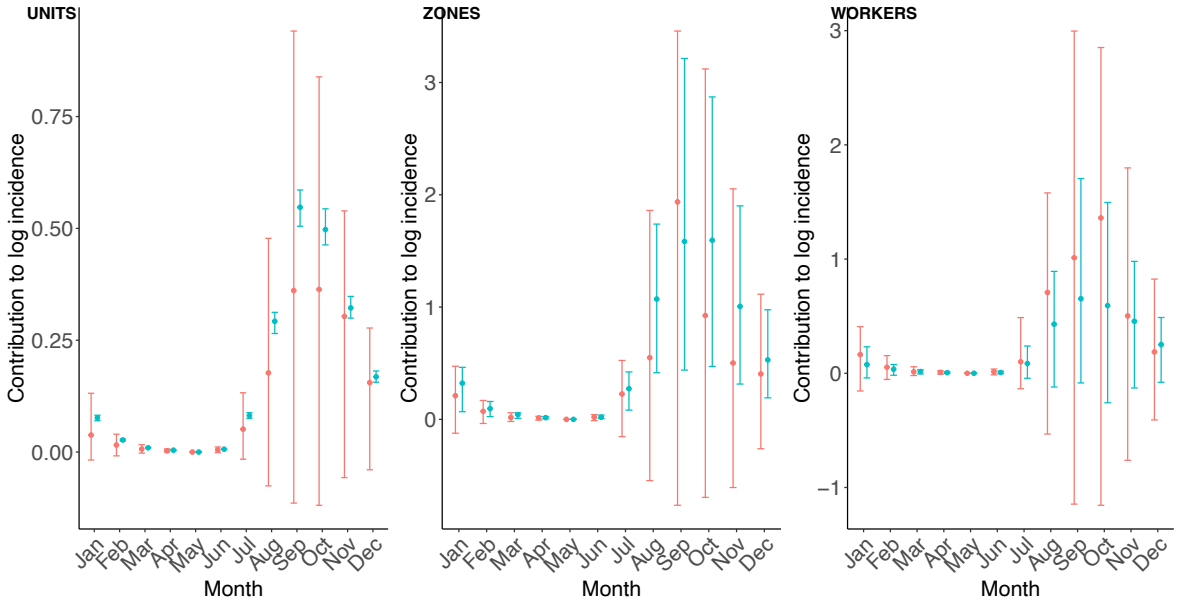

Figure S9. Posterior distributions of monthly random effects. Effect of climate on the annual cycle of malaria in Surat. The difference in the monthly random effect marginal posterior distributions for models with just the random effects (in red) and those including the climate factors, temperature and humidity, (in blue) at the three evaluated levels: (A) *Units* (B) Zones, and (C) Workers.

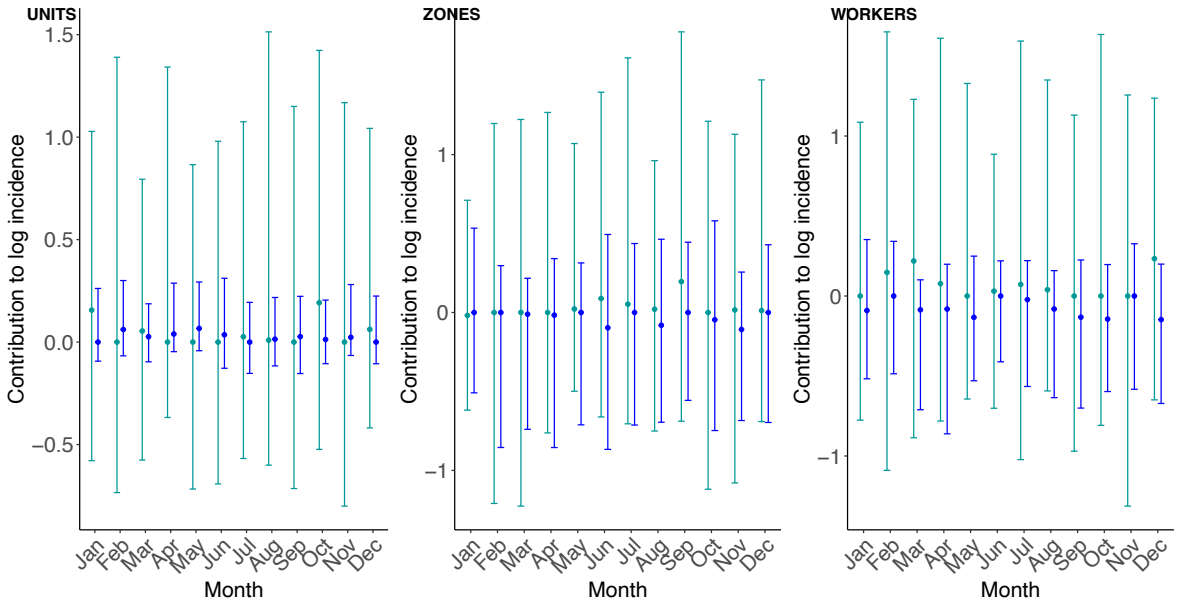

Figure S10. Posterior distributions of annual random effect. Interannual random effects for malaria in Surat. The difference in the interannual random effect marginal posterior distributions for a baseline model (in green) and the climate model (in blue).

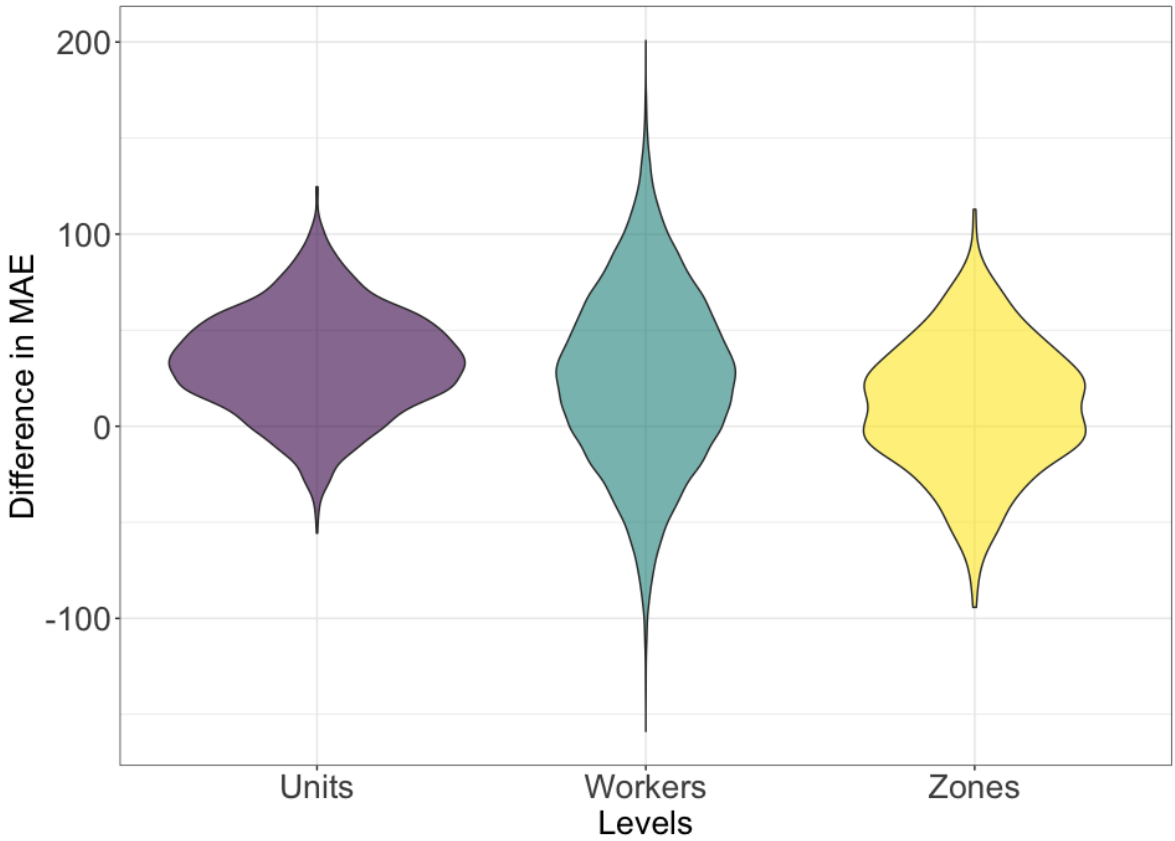

Figure S10. Predictive ability of the models. Each violin plot shows the distribution of **the differences between** administrative units with positive values (MAEbaselinemodel–MAEbestmodel>0), indicating that accounting for climate and socioeconomic variability improves the estimation of malaria incidence rates in these places, as its inclusion results in a smaller difference between the modeled values and the observations, than using the seasonal spatial model alone.

Appendix 2

Data description and processing

Socioeconomic data

A database of candidate drivers of urban malaria risk in Surat was generated for each reporting unit (zones, units, and workers) for 2010 (the year of the most recent census) using an ordinary kriging method to generate an estimated interpolation surface for each covariate (see below). For this, we used census data from the District Census Handbook at the district level for 2010 from the Directorate of Census Operations, Gujarat. We selected variables with a significant effect on malaria transmission, according to the literature[1,2] (see S1 table 1). In addition, population density and slum density were calculated using annual population estimates from the Surat Municipal Corporation (SMC). Data on urbanization, housing, water provision, water storage, sanitation, and water scarcity were obtained from water management surveys conducted by Taru Leading Edge, a leading development advisory company in India (http://taru.co.in/). This household survey included 90 locations and 400 households (S1 Fig 8), selected to be uniformly distributed throughout the city. These interpolated covariates then became the input for our PCA analysis in the main text.

Figure S1.8. Spatial distribution of the survey points in the city of Surat.

Climate data:

We obtained 8-day Land Surface Temperature (LST) from both Terra and Aqua satellites of MODIS between 2008 and 2015 to estimate monthly surface relative humidity within the city for the same period we had malaria cases. First, we used daytime and night-time 8-day composite LST data from MODIS at 1 km spatial resolution (S1 Fig 9). These datasets were downloaded from the NASA server at (https://earthdata.nasa.gov/eosdis/daacs/laads).

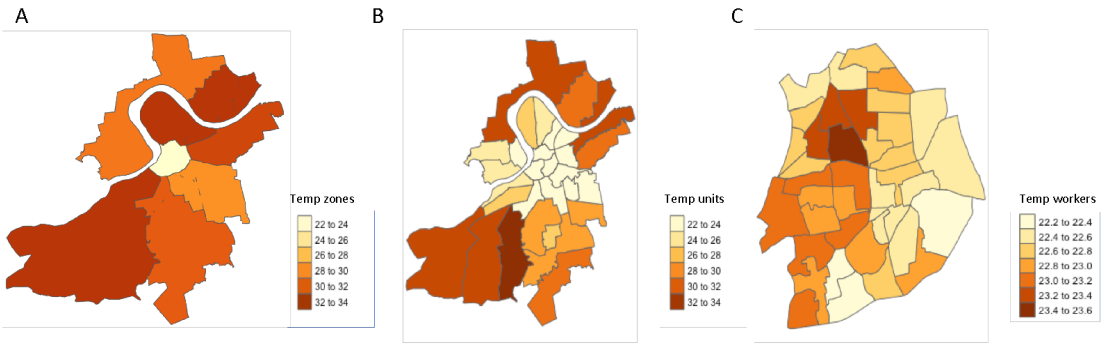

Figure S1 9. Variation in the temperature (Land Surface Temperature) extracted from MODIS at different levels of aggregation, A) Zones, B) Units, and C) Workers' Units. For the highest resolution, given a large number of units, only a part of the map is shown here, specifically for the worker units comprising the three units at the city's core.

Although satellite-derived LSTs offer the advantage of covering a larger area than the local stations, these measurements may experience quality limitations mainly due to cloud contamination. Thus, to evaluate the quality of these products, we compared the satellite data to local measurements from ten temporary meteorological stations established throughout the city of Surat by TARU. We compared the monthly temperature and humidity observations with the satellite-derived measures for 2014. We specifically calculated the Pearson's correlation coefficient between the interpolated satellite-derived data and the meteorological station data, obtaining an R2 = 0.72 for RH, which indicates a high linear correlation between the two data sets.

1. *Relative humidity from satellite products*

Relative humidity, RH, reflects the atmosphere's moisture. To estimate the relative humidity in the city of Surat, we combined precipitable water (PW) datasets for the city from the product MOD07 from MODIS. Namely, we first estimated specific humidity as a function of PW and PW1 using the expression derived by Liu (1984): [3]

$$Q=0.001\times(-0.0762PW1 + 1.753\ PW + 12.405) \quad (1).$$

Then we used the SRTM digital elevation data produced by NASA to retrieve elevation values and calculate the vapor pressure using the statistical relationship presented in Peng et al. 2006[4].

$$Pa = 1013.3-0.1038H \quad (2).$$

Finally, we used data from MODIS Level-2 to extract surface air temperature. These data are derived from known emissivity at different spectrum bands, as observed from the satellites in clear sky conditions. With these three quantities, we were able to calculate values of RH using the expressions from Wang 1987[4], where RH is the ratio of vapor pressure (e) and saturation vapor pressure (es):

$$RH = \frac{e}{e_s} * 100 \quad (3)$$

Vapor pressure e depends on air pressure (Pa) and specific humidity (Q) as given by $e = Q + P_a/0.622$, and saturation vapor pressure (es) is given by $e_s = 611e(17.27 * T_a/237.3 + T_a)$ where T_a represents the air temperature from MODIS .

Kriging description

Kriging is a technique that generates an estimated interpolation surface from a set of data points [5]. This technique incorporates inference from the spatial structure of data points to derive estimations at unmeasured locations. Here we used universal kriging, which incorporates spatial trends into the interpolation process. In mathematical terms, the random function Z(x), representing the variable of interest, is split into a deterministic drift m(x) and a random function e'(x) with zero means.

$$Z(x_0) = m(x) + e'(x)$$

where m(x) is a structural component associated with a constant mean value or a constant trend and $e'(x)$ a second-order stationary random function spatially correlated component, known as the variation of the regionalized variable. In universal kriging, the mean is a function of the site coordinates. Then, m(x) follows the equation:

$$m(x_0) = \sum_{i}^{n} \propto_i p_i$$

where $\propto_i$ is the local trend or drift coefficients and p_i is a function of the site coordinates. We assume that the drift is a smooth function of the coordinate vector x and that it can be represented by a linear function $m(x) = \propto_0 + \propto_1 x + \propto_2 y$ where x and y are the coordinates of the *ith* control point and $\propto$ are the drift coefficients.

2. Data collating and harmonizing (interpolation process)

Given that we are using datasets from multiple sources and multiple levels of aggregation, we had to collate, integrate, and harmonize data across different resolution scales to perform our analysis. For the purpose of interpolation, we relied on ordinary kriging (OK). We followed the protocol described below for our climate and socioeconomic data:

a) One assumption that needs to be met in ordinary kriging is that the mean and the variance of the studied entity are constant across the study area. Thus, we removed any global spatial trend in our data.

b) For this form of kriging, we calculated the *experimental variogram* to see the spatial pattern of temperature, relative humidity, and socioeconomic data. More specifically, we are interested in how these attribute values vary as the distance between location point pairs increases.

c) Then, we fitted a mathematical model to the *experimental variogram*. Specifically, we were looking for a function that describes the degree of spatial dependence and continuity across data. We generated two theoretical models to approximate the data best. First, we used a theoretical semi-variogram using a spherical model fit, and then we fitted an exponential model for the semi-variogram.

d) Finally, we generated a grid of 250mx250m for the Surat area. With the appropriate variogram model selected, we used kriging to predict the respective surfaces for temperature, relative humidity, and socioeconomic variables. The final model uses the measure of variability between points at various distances. Points nearby likely exhibit more similar values, and as the distance between points increases, the similarity between values most likely decreases. In this application, we assume that temperature and socioeconomic measurements that are further apart will vary more than those closer together. Using the spherical model fit and plot, we used the Krige function in R to krige the data.

Appendix text 3

Bayesian modeling

A negative binomial model was implemented to allow for over-dispersion in urban malaria count data[6]. The formulated model account for temporal effects (seasonal and interannual) as well as

spatially explicit effect, which incorporated: (i) a spatially unstructured random effect of introducing an additional source of variability/over-dispersion in space (a latent effect) and (ii) a spatially structured random effect to explicitly account for spatial autocorrelation and in so doing, weight relative risk in a region according to the relative risks in neighboring regions. To infer the coefficient of the regression, we used an integrated nested Laplace approximation for approximate Bayesian inference[7]. The integrated nested Laplace approximation (INLA) is a method for approximate Bayesian inference. In the last years, it has established itself as an alternative to other methods, such as Markov chain Monte Carlo because of its speed and ease of use via the R environment. Although the INLA methodology focuses on models that can be expressed as latent Gaussian Markov random fields (GMRF), this encompasses a large family of models used in practice[8,9].

In our models, the parameters were estimated using the R-INLA package, the Integrated Nested Laplace Approximation (INLA, www.r-inla.org) in R. Parameter uncertainty is accounted for by assigning prior distributions to the parameters. First, the random seasonal effect of accounting for the seasonal autocorrelation was incorporated as a month effect $m_t = 1, \ldots.12$ (Jan to Dec), this effects were assigned as a cyclic random walk of order one $m_{t(t)} - m_{t(t-1)} \sim N(0, var)$ where m_1 represents the parameter estimate for January. Regarding the interannual variation, we incorporated an exchangeable random effect specified for each year ($a_t \; where \; t' = 1 \ldots.7$, 2008–2014) to allow for additional sources of variability that the model covariates could not capture. For both random effects, we assigned a Gaussian distribution with zero mean and large variance for the prior. Also, we assigned the R-INLA default hyperparameter to the precision ($t = 1/\sigma^2$), that is, t ~Gamma(a, b) with shape parameter a = 1 and inverse-scale parameter b =1×10^{-5}. Finally, we used the modified Besag-York-Mollie (BYM) model to account for spatial correlation in malaria incidence. As for the classical BYM this modification incorporates an unstructured spatial random effect to account for independent area-specific noise and over dispersion, included here via ϕ_i, and a structured effect incorporated as a conditional autoregressive model to allow for spatial dependency between neighboring areas via v_i (with a conditional prior referred as CAR)[10,11]. To incorporate potential nonlinear associations in the climate covariates, we used nonlinear exposure-lag functions f.w(Xist, l) of temperature, x1st, and the humidity, x2st, with lags, l, from zero to 3 months. We further compared models with temperature and relative humidity that were either spatially resolved or averaged over all the spatial units at the global level of the city. This allows us to examine whether climate factors better explain the malaria data by influences at the local vs. regional level.

Model comparison and evaluation of goodness-of-fit were performed at each of the levels, for all the models were assessed their fit using the deviance information criterion (DIC) [12], the Watanabe-Akaike information (wAIC) criterion[13], and a R^2_{LR} statistic for mixed effects models, based on a likelihood ratio (LR) test between the candidate model and an intercept only (null) model[14,15] , for cases aggregated into five discrete levels. Smaller values of DIC and wAIC indicate a better-fitting model. The likelihood ratio R^2_{LR} ranges from zero to 1, with one corresponding to a perfect fit for any reasonable model specification[16]. We checked the convergence of the individual parameter estimates for each resulting model and calculated the potential scale reduction. (This reduction is the ratio between the within-chain and posterior variance estimate variances. Values below 1.1 are considered acceptable in most cases; see [15] Gelman et al. 2004 for details). To allow for uncertainty in the response variable (i.e., malaria cases), given the model parameters, posterior predictive distributions of the response variable were simulated by drawing random values from a negative binomial distribution with mean and scale parameters estimated using 10.000 samples from the posterior distribution of the parameters in the model and computing the median number of malaria cases from the simulations. Finally, model selection was performed separately for the data at each spatial scale: the intermediate spatial resolution (units level, 32 units) and at coarser (zone level, seven units) and finer (worker unit, 486 units) spatial resolutions, respectively.

References

1 Lana RM, Riback TIS, Lima TFM, *et al.* Socioeconomic and demographic characterization of an endemic malaria region in Brazil by multiple correspondence analysis. *Malar J* 2017; **16**: 397.

2 Santos-Vega M, Bouma MJ, Kohli V, Pascual M. Population Density, Climate Variables and Poverty Synergistically Structure Spatial Risk in Urban Malaria in India. *PLoS Negl Trop Dis* 2016; **10**: e0005155.

3 Liu WT, Niiler PP. Determination of Monthly Mean Humidity in the Atmospheric Surface Layer over Oceans from Satellite Data. *J Phys Oceanogr* 1984; **14**: 1451–7.

4 Peng G, Li J, Chen Y, Norizan AP, Tay L. High-resolution surface relative humidity computation using MODIS image in Peninsular Malaysia. *Chin Geogr Sci* 2006; **16**: 260–4.

5 Oliver MA, Webster R. Kriging: a method of interpolation for geographical information systems. *Int J Geogr Inf Syst* 1990; **4**: 313–32.

6 Longini IM, Koopman JS, Haber M, Cotsonis GA. STATISTICAL INFERENCE FOR INFECTIOUS DISEASES. *Am J Epidemiol* 1988; **128**: 845–59.

7 Rue H, Martino S, Chopin N. Approximate Bayesian inference for latent Gaussian models by using integrated nested Laplace approximations. *J R Stat Soc Ser B Stat Methodol* 2009; **71**: 319–92.

8 Moraga P, Dean C, Inoue J, Morawiecki P, Noureen SR, Wang F. Bayesian spatial modelling of geostatistical data using INLA and SPDE methods: A case study predicting malaria risk in Mozambique. *Spat Spatio-Temporal Epidemiol* 2021; **39**: 100440.

9 Rue H, Riebler A, Sørbye SH, Illian JB, Simpson DP, Lindgren FK. Bayesian Computing with INLA: A Review. *Annu Rev Stat Its Appl* 2017; **4**: 395–421.

10 Fletcher IK, Stewart-Ibarra AM, Sippy R, *et al.* The Relative Role of Climate Variation and Control Interventions on Malaria Elimination Efforts in El Oro, Ecuador: A Modeling Study. *Front Environ Sci* 2020; **8**: 135.

11 Grillet ME, Moreno JE, Hernández-Villena JV, *et al.* Malaria in Southern Venezuela: The hottest hotspot in Latin America. *PLoS Negl Trop Dis* 2021; **15**: e0008211.

12 Watanabe S. Equations of states in singular statistical estimation. *Neural Netw* 2010; **23**: 20–34.

13 Kramer M. R^2 STATISTICS FOR MIXED MODELS. *Conf Appl Stat Agric* 2005; published online April 24. DOI:10.4148/2475-7772.1142.

14 Magee L. R 2 Measures Based on Wald and Likelihood Ratio Joint Significance Tests. *Am Stat* 1990; **44**: 250.

15 Gelman A, Rubin DB. Inference from Iterative Simulation Using Multiple Sequences. *Stat Sci* 1992; **7**. DOI:10.1214/ss/1177011136.

16 Gelman A, Carlin JB, Stern HS, Dunson DB, Vehtari A, Rubin DB. Bayesian Data Analysis, 0 edn. Chapman and Hall/CRC, 2013 DOI:10.1201/b16018.